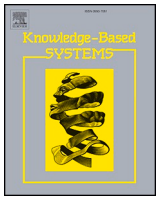
Knowledge-Based Systems 299 (2024) 112088

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Machine learning Ethereum cryptocurrency prediction and knowledge-based investment strategies

Adrián Viéitez [a], Matilde Santos [b,*], Rodrigo Naranjo [a]

[a] Dept. Computer Architecture and Automatic Control, Computer Sciences Faculty, University Complutense of Madrid, 28040 Madrid, Spain
[b] Institute of Knowledge Technology, University Complutense of Madrid, Computer Sciences Faculty, 28040 Madrid, Spain

ARTICLE INFO

Key words:
Machine learning
Knowledge-based systems
Neural networks
Cryptocurrency market
Ethereum
Forecasting
Investment decision system

ABSTRACT

This work proposes a novel methodology to help in decision making in the cryptocurrency market. Two investment strategies have been designed for Ethereum (ETH), based on predictions of the price and trend of this cryptocurrency using real data. The two Ethereum cryptocurrency prediction systems rely solely on past values of other contextual stock indices, market indicators and online trends, and ignore any technical indicators of price evolution. Real data from cryptocurrency market has been collected and processed with different feature selection methods. Applying a regression approach with Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) networks, prediction models for the ETH price for 1, 7 and 15 days are obtained and compared. Also, support vector machine (SVM) is applied to predict the ETH price trend by applying a classification approach. In both approaches, sentiment analysis has been included to check its effect on the prediction results. The reliability of these prediction models in the current market has been evaluated by designing two original knowledge-based investment strategies. They are tested over two different time periods with real cryptocurrency market data. The results show that it is possible to generate up to 5.16 profit factor with few operations using these models. Furthermore, adding sentiment analysis has shown to have little influence. In this way, we contribute to the advancement of our knowledge of this volatile and still young cryptocurrency market, and specifically of the evolution of Ethereum and the factors that can influence its behavior.

1. Introduction

With the blockchain serving as its foundation, cryptocurrencies (CC) are a type of digital exchange that guarantees that all transactions are carried out through a strong encryption procedure. Blockchain is made up of blocks with an expanding list of records connected by encryption to prevent data manipulation. Every block in the blockchain contains transaction data, a new timestamp, and the cryptographic hash value of the previous block. The majority of electronic currencies are produced by mining, an incentive-based process that creates new units and verifies transactions while also adding them to the core of already-existing ones [1].

Accurately forecasting the evolution of the prices of stocks and other financial assets has always been the ambition of those who make investments their job, a way to obtain additional income or simply to find long-term solutions to save money in a more productive way than leaving it in a savings account. In this sense, the cryptocurrency market has become another investment field of great interest and growth for more and more people and companies.

There are many challenges linked to this recent market, such as the lack of large amounts of historical data, market volatility, unknown factors that could influence its evolution, and its development on online platforms, which offers conditions never seen before, such as 24/7 trading windows, easy access to millions of people and the information and misinformation campaigns to which this entire public is exposed.

In addition, some problems have arisen in dealing with these markets that are so loosely linked to traditional markets. They are very volatile markets, which experience rapid changes, and the large amount of data they handle comes from different sources. Moreover, in this new era of social networks, alternative information that is very different from raw data, such as sentiment analysis, can be even included. In fact, most work on this specific market shows a lack of use of contextual data that could help forecasting.

This work aims to estimate the price and trend evolution of a particular cryptocurrency, Ethereum (ETH), by using machine learning and artificial intelligence techniques to address some of these problems.

* Corresponding author.
E-mail address: msantos@ucm.es (M. Santos).

https://doi.org/10.1016/j.knosys.2024.112088
Received 11 March 2024; Received in revised form 27 May 2024; Accepted 6 June 2024
Available online 12 June 2024

Indeed, the prediction systems are based only on past values of other contextual stock indices, market indicators, and online trends, and ignores any technical indicators in favour of the price evolution itself. Furthermore, in view of the results of the forecast and the lack of intelligent investment methods in the cryptocurrency market found in the literature, we propose two knowledge-based investment strategies that can help obtain financial benefits in such a complex and volatile environment, full of opportunities but also high risks.

As said, in this study the chosen cryptocurrency is Ethereum. Although most works use Bitcoin (BTC), in this case it has been considered more interesting to use Bitcoin as another input feature than as an object of study itself. This allows the possibility of better predictions due to the influence that Bitcoin price has on other cryptocurrencies, also known as altcoins, and more specifically on Ethereum [2].

The following are this work's primary and original contributions:

- The procedure from the selection of data, variables and market indicators that can be relevant to the application of ML techniques for forecasting in the CC market is developed and described.
- Granularity, time frame, and feature selection analysis of cryptocurrency market data, traditional stock market indicators and online statistics, with different techniques.
- Ethereum price prediction using machine learning techniques, specifically Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) networks, designed for this purpose under a regression approach.
- An intelligent classification approach to forecast Ethereum price trend using support vector machine (SVM).
- Two novel investment strategies have been designed. These knowledge-based decision systems allow, on one hand, to validate the forecasting models in a real time period, and on the other hand to show how profitable these intelligent decision systems can be.
- Development of additional functions that can extract sentiment analysis from two different online platforms to enrich the information and verify the usefulness of including it in prediction and investment strategies.

This study contributes to existing state of art in several ways. First, we have worked with ETH, a less common CC than the more widely used BTC in most academic works, since the latter has a longer life and is therefore better known. This entails the difficulty of working with less data, also knowing that the first launch period of the ETH cryptocurrency, more recent than BTC, is noisier. For this reason, greater effort has been required in the configuration of the techniques used for prediction, as well as in the preprocessing of the information.

Furthermore, there are few works that refer to feature selection and the vast majority of those that explore this path do not provide any contribution outside the world of cryptocurrencies. In the current context of a market increasingly related to the traditional stock market and its variables, this work offers a very innovative and extensive approach to select features from three different fields (traditional stock market, online statistics and cryptocurrency market), applying and comparing different feature selection techniques for different forecast horizons. Even more, the results have been compared with systems that incorporated additional functions to introduce sentiment analysis, corroborating results obtained in other works. However, in our case instead of using Twitter, sentiment from the much less frequent Reddit and Google news has been extracted. Unlike previous studies, the prediction models have been validated with novel investment strategies that have allowed their efficiency to be evaluated. This allows us to offer a general computational methodology, which can be applied to any type of cryptocurrency. Indeed, the predictions on which the investment decision systems are based have been made without taking into account any technical indicator extracted from the price evolution itself, based solely on the values of other contextual stock indices, market indicators and trends in line. This facilitates its application to other cryptocurrencies.

Finally, this study has allowed us to advance our knowledge of Ethereum's price behaviors in particular, and to have a closer view of the cryptocurrency market in general. This is a young market, with little data and that is very volatile. Methodologies based on ML have been found that have proven valid for dealing with this type of information and obtaining useful and practical results.

The structure of the article is the following. Section 2 discusses some related scientific papers. In Section 3, the database and the techniques used are presented. Section 4 explains the ETH price forecasting with the neural network regression approach and Section 5 describes the ETH price trend prediction with the classification approach. The investment strategies based on those approaches are developed in Section 6. Conclusions and future research lines end the manuscript.

2. Background

2.1. Ethereum and cryptocurrency market

Cryptocurrencies are digital representations of values that are meant to function as a peer-to-peer ("P2P") substitute for legally recognized money issued by governments, according to policy makers. It can be changed into legal tender and vice versa, it is also used as a general-purpose means of trade (unaffected by central banks), and it is safeguarded using a system called cryptography [3].

Ethereum, [4], launched in 2015, aims at creating a universal blockchain-based application platform. It addresses several limitations of the Bitcoin scripting language, incorporating a Turing complete language, which means that Ethereum allows all kinds of calculations, including loops, to overcome many shortcomings of the other cryptocurrencies programming language.

Because of its abstract layer, anyone can design their own ownership rules, transaction formats, and state transition mechanisms. Smart contracts, a collection of cryptographic guidelines that only take action when specific requirements are satisfied, are used to achieve this [4]. The Ethereum network's consensus relies on a modified version of the Ghost protocol, often known as the Greedy Heaviest Observed Subtree [5]. It was developed to address the network's stale block problem. Stale blocks can happen when a mining pool's aggregate processing power is greater than that of the other groups of miners. This means that the blocks from the first pool will contribute more to the network, which will lead to a centralization problem. These dead blocks are taken into account by Ghost Protocol when determining the longest chain. The solution to the centralization issue is to give stales block rewards. In other words, even if a miner's block (referred to as an uncle) does not make it into the main blockchain, they are still compensated. Up to seven generations of uncles are included in Ethereum's adaptation of the Ghost protocol [6].

An environment where investors trade assets to make money is called a financial market. Over time, asset prices fluctuate. Therefore, by anticipating these shifts, investors can buy assets that might appreciate in value and turn a profit. There is at least one asset in every market that is thought to maintain its value throughout time. This unique asset, which may be used to buy any asset on the market, serves as the benchmark for measuring the prices of all other assets. That's why it is called cash. The US dollar (USD) is the most widely used cash asset in the stock and fiat currency markets. Conversely, the most well-liked cryptocurrency in markets is Tether (USDT). This is so because the USDT's exchange rate to the USD was intended to remain constant. As a result, cryptocurrency investors can quickly exchange their holdings for this risk-free asset if they wish to close their positions or end their trading sessions.

The designated hours during working days during which investors are permitted to trade equities are known as trading sessions in stock exchanges [7]. But since bitcoin markets are operating

around-the-clock, they are not subject to this restriction [8]. However, we refer to the period of time during which the implemented algorithms are permitted to conduct transactions in a simulated setting as a trading session. With that said, it is widely stated through all cryptocurrency investment platforms that the opening time of this virtual trading sessions for cryptocurrency is the 00:00 UTC and the closing time the 23:59 UTC.

In both, cryptocurrency markets and stock markets, it is common to see the sessions represented as candlesticks, and indeed it is a good way to offer an overview of a session and its evolution [9]. The candlesticks do not necessarily represent sessions, since usually they can be adjusted to represent different time frames, going from minutes, to hours, days, weeks, etc. The main characteristics of a candlestick, which can be "bullish" (white body) or "bearish" (black body), are the following.

- **Open price:** the first price at which a cryptocurrency is traded after the opening time of the session.
- **Close price:** price at which has been ultimately traded before the closing time deadline.
- **High price:** highest price at which the cryptocurrency has been traded during the season.
- **Low price:** lowest price at which the cryptocurrency has been traded during the season.
- **Adjusted close:** in stock market, it amends the close price considering other factors like dividends, stock splits, or offerings of new stocks. In the Cryptocurrency market it does not reflect any changes with respect to the close price.

That is, the increasing candles correspond to sessions where the price at the end is higher than at the beginning, and decreasing candles mean the opposite.

2.2. Related works

Machine learning techniques, and Artificial neural networks (ANN) among them, have been used for predicting in a wide range of fields. Some authors consider ANNs are superior to statistical techniques in forecasting because they have a number of attractive properties. For example, [10] demonstrated that ANNs perform comparably better with long prediction horizons when compared to the Box-Jenkins approach for forecasting. When predicting the consumer price index and the anticipated number of cancer patients in a Yemeni province, [11] comes to the same conclusion. Up to five DL NN models are integrated with an improved ARIMA in [12] to anticipate monthly crude oil prices. In any case, to adjust artificial neural networks (ANNs) for maximum accuracy, it is imperative to identify the optimal network topology and training protocol for a given issue. The choice of input variables for modeling and prediction is a crucial issue as well [13].

Over the years, ANNs have been also applied for stock market prediction, proving that they are more effective than regression methods. For instance, it has been shown that the error in the prediction of the price-change magnitude using MLP neural network is smaller than using linear regression [14]. Moreover, when the feedforward MLP neural network predicts the direction of the changes correctly, the change magnitude is very close to the real one in comparison with linear regression [15]. This conclusion not only shows the convenience of using neural networks over regression methods, but also points towards recurrent neural networks. In [16] ARIMA and LSTM are compared regarding accuracy, as representative techniques when forecasting time series data, and the conclusion is that LSTM-based algorithms improved the prediction by 85% on average compared to ARIMA.

Specifically on the Cryptocurrency market and price prediction, some recent works can be found in the literature as this topic has lately gained more interest. Most of them study the performance of a few cryptocurrencies, primarily Bitcoin, since there is a greater amount of historical data available than for other alternative cryptocurrencies. For this Bitcoin market, investment strategies range from a set of technical trading rules, as in the case presented in [17], to more sophisticated ones. Some notable exceptions analyze various cryptocurrencies, as in [18], where it is shown that Bitcoin, Litecoin and Ethereum markets show a clear tendency that evolves from less to more efficiency, in contrast to Ripple, Stellar and Monero, with changes that alternate periods of efficiency and inefficient.

Some studies compare the performance of different Machine Learning strategies, such as ARIMA, SVR, RF-Regressor, XGBoost, MLP, LSTM, GRU, CNN, or their combinations, LSTM+CNN, GRU+CNN. The worst performance attending to RMSE and R-squared measurements is obtained with the ARIMA method and the best results are usually given by Recurrent Neural Networks [19] when the forecast is stated as a regression problem, while support vector machines have demonstrated to be the strongest classifier [20]. For example, in [21], the prediction of copper spot prices from the New York Commodity Exchange is examined through the use of support vector regression (SVR), a machine learning technique, in conjunction with various model schemas (recursive, direct, and hybrid multi-step). The hybrid direct-recursive performs better, according to the results. Using extracted and unique information, the study by [22] uses SVM and ANN to accurately predict the value of the Bitcoin price. Using daily historical data, other research comparing the performance of several RNNs conclude that LSTM and GRU perform quite similarly when forecasting the prices of Bitcoin, Ethereum, Cardano, Tether, and Binance [23]. The paper by [24] presents a very complete overview in this field, concluding that LSTM networks are the ones offering the best performance for regression of Bitcoin price.

A study published in [25] came to a similar conclusion by using a long short-term memory algorithm to predict the values of four different cryptocurrency types: Ethereum, AMP, XRP and Electro-Optical System. In order to achieve the primary goal of lowering the financial risks involved in conducting electronic business, a deep neural network algorithm was improved in [26] to forecast the price of Bitcoin. Good prediction was accomplished by using useful data, such as transactions and currency returns. In order to forecast the price of the Bitcoin cryptocurrency, the work by [27] compares the GRU and MLP models. It finds that the MLP model produced extremely efficient regression.

Regarding the features that should be considered in the prediction, some works mention their impact. For instance, [28] explores several financial features and shows once again how RNN as GRU and LSTM outperform traditional machine learning models but brings up the importance of introducing dropout in the layers to reduce overfitting and improve performance. The work presented in [29] makes use of Soft Gradient Boosting Machine and Rain Forest Tree to predict the prices of some cryptocurrencies such as Bitcoin, Ripple and Ethereum, and they confirm that the inclusion of additional features into the model led to an increased from 1 to 3% accuracy on average. Additionally, this work shows that training datasets with large windows of past values (28 days) help have a better trained model. Similarly, [30] demonstrates that LSTM networks work well using 30 days of previous values for one day ahead prediction, but the performance improves using fewer past values since, due to the design of the model, it frequently depends more on short-term dynamics. Nevertheless, this might be different applying a large number of input features, what was not tested in the mentioned work. The importance of feature selection is also discussed in the work [31], which highlights the forecasting power of macroeconomic factors and technical indicators, finding that the macroeconomic factors present a better capacity to forecast Bitcoin.

About the forecast horizon, it has been observed that in the very short-term prediction -between 15 and 60 min- of changes in the price of Bitcoin as a binary variable, the accuracy obtained is around 50–55%, improving as the prediction horizon increases [32 It is not a good result and seems to suggest that it would be better to focus on short-medium term, at least one-day predictions.

In [33] the authors propose a non-parametric solution for predicting day-ahead crypto-currency prices. The method is called deep state-space

model as it combines the probabilistic formulation of the state-space model with the estimation of the function approximation ability of the deep neural network. The influence of the window size and depth is analysed. Results are compared with different machine learning models. Other works [34] point towards the same forecast horizon but from a different approach, which despite not being able to find a good solution for predicting the price of Bitcoin, do yield some interesting conclusions that show that the performance of RNN decreases when the prediction horizon increases to more than one day.

Going one step further, some studies try to understand the connection of global social and economic events in the evolution of the price of Cryptocurrencies. For example, [35] shows that the cryptocurrency market actively reacts to economic policy uncertainty and geopolitical risks. Following these principles, some works study the possibility of extracting sentiment data from online websites, forums, and social networks, and using it as a feature to obtain better prediction results. In [36], a comparison is made between using or not using sentiment data in the price forecasting of Bitcoin, Ethereum and Ripple, using three different machine learning techniques: MLP, SVM and RF. Their results do not draw a clear conclusion; in some cases Twitter data helps to slightly improve the prediction, even the prediction can be made using only the sentiment data, but in other scenarios including the sentiment analysis worsens the performance of the model. This may be because the sentiment data comes only from Twitter and may not be well filtered, suggesting that better pre-processing of the data could help improve the results. Other works that yield interesting results in this sense are, for example, the one by Bouri and Gupta [37], where the newspaper-based measure and an internet search-based measure of uncertainty in predicting Bitcoin returns are considered. Also the work by [38], shows that the predictability patterns of intraday returns of Bitcoin change in the presence of large intraday price jumps, the publication of FOMC announcements, liquidity levels and the outbreak of COVID-19. Furthermore, this study concludes that intraday reversal, which is unique to the cryptocurrency market, may be related to investors' overreaction to non-fundamental information and overconfidence bias.

Other works, such as [39], implement robust sentiment analysis based on cryptocurrency-specific lexicon, in combination with bivariate Granger causality tests. It is shown that information from Twitter can be used to predict the prices of some of the top ten cryptocurrencies. But in general, there are no works that show a clear benefit when using sentiment analysis, although most of them consider that it is the right direction to investigate and take the prediction to the next level of precision.

After studying the related literature, it can be observed that, unlike the works mentioned, the intelligent system proposed here not only predicts with greater accuracy the price and price trend of Ethereum, a cryptocurrency little considered in the literature, but also proposes a strategy investment that has been proven successful in generating profits.

The main difference with the works mentioned here is that in this proposal an attempt has been made to integrate different knowledge and make it more complete. As has been mentioned, some works use fewer characteristics, do not combine as many indicators or do not include sentiment analysis. Others only apply one type of NN or an approximation, and do not work with different approaches such as regression and classification. Most do not propose investment strategies. And the fact of showing the entire process and comparison of various tools for the different phases of cryptocurrency prediction and management will not be reflected in other papers, and even less so for ETH.

3. Materials and methods

3.1. Methodologies applied

Based on the scientific literature and the properties of some intelligent techniques, we have chosen some Machine Learning (ML) methods for predicting ETH. Machine learning is a subset of Artificial Intelligence techniques and algorithms that, using data, automatically improve through training. The diversity of techniques and principles it comprises make this branch of artificial intelligence very flexible and with a high capacity for adaptation and computation.

Among them, neural networks stand out, which are capable of learning any non-linear function, no matter how complex it may be, which is why they are called universal approximators. Hence its usefulness for identification and prediction tasks.

A typical neural network (NN) is made up of several neurons, which can be distributed in layers. Each neuron is a local processor with a high degree of connectionism with the rest of the neurons, which when activated generate an output value. The weighted connections of previously active neurons in turn activate other neurons, while sensors that detect their environment activate input neurons. Finding weights that make the NN behave in the desired way is the goal of the learning process.

NNs can have different architectures. We can distinguish feedforward NNs, which are the simplest (an example is the multilayer perceptron, MLP); convolutional neural networks (CNN), which apply filters to information to process it and obtain new data, which usually have a spatial component, and recurrent neural networks (RNN), which are usually applied to temporal and sequential data. NN, and specifically RNN, outperforms linear regression in stock market prediction[14,15].

In this work we have used some type of RNN, particularly:

- Long Short-Term Memory (LSTM) neural networks, which can store information and use it again shortly after. They are well-liked for handling sequential data, including time-series data, because they can pick up long-term dependencies. They have been used here because they have been proved efficient in stock price prediction, including cryptocurrencies [19,25]. They are able to find patterns and forecast future prices based on a sequence of stock prices throughout time. Even more, the vanishing gradient problem is handled using LSTMs. An input gate, an output gate (a straightforward MLP), and a forget gate are utilized by the long-short-term memory cell. Data is gathered for processing in the subsequent stage using the cell state and concealed state. The equations of the gates are:

  Input Gate

$$i_t = (W_i h_{t-1} + W_i h_t) \tag{1}$$

  Forget Gate

$$f_t = \left(W_f h_{t-1} + W_f h_t\right) \tag{2}$$

  Output Gate

$$o_t = (W_o h_{t-1} + W_o h_t) \tag{3}$$

  Intermediate Cell State

$$C = \tanh(W_c h_{t-1} + W_c h_t) \tag{4}$$

  Cell State (next memory input)

$$c_t = (i_t * C) + (f_t * c_{t-1}) \tag{5}$$

  New State

$$h_t = o_t * \tanh(c_t) \tag{6}$$

  Where the input is denoted by X_t, the output is h_t, the parameter matrices are W,U, and the vector is f.

- The GRU, or Gated Recurrent Unit network, is appropriate for forecasting patterns in the stock market since it performs better at handling non-linear situations than conventional machine learning techniques. Additionally, GRU may outperform other RNN such as

LSTM due to its superior performance with fewer parameters, which makes it simpler and faster, but also less flexible and powerful. It has also been successfully used for cryptocurrency prediction [23]. It has an update gate (z) and a reset gate (r), but no output gate. These gates determine what data should be sent to the output by using vectors. How to merge the new input with the prior memory is specified by the reset gate. The following equations are part of the GRU:

Update gate

$$z = \sigma(W_z h_{t-1} + U_z x_t) \tag{7}$$

Reset gate

$$r = \sigma(W_r h_{t-1} + U_r x_t) \tag{8}$$

Cell state

$$c = \tanh(W_c(h_{t-1} * r) + U_c x_t) \tag{9}$$

New state

$$h_t = (z * c) + ((1 - x) * h_{t-1}) \tag{10}$$

Support Vector Machines (SVM) have also been applied in this study. SVM deals with sparse solutions, referring to the fact that SVMs only employ a portion of the training data in their prediction processes. As a result, the algorithm becomes less prone to overfitting and more efficient. They have been shown able to accurate predict the value of the Bitcoin price, for instance [22]. By using kernel functions to project the input space into a higher dimensional characteristics space, it provides a nonlinear and solid solution. To distinguish examples from the two classes, they transfer the input (x) into a high-dimensional feature space ($z = \varphi(x)$) and build an ideal hyperplane described by $w \cdot z - b = 0$. This is accomplished for SVMs with L1 soft-margin formulation by resolving the primal problem.

$$\min \frac{1}{2} w^2 + C \sum \xi_i$$

$$\text{s.t. } y_i(w \cdot z - b) \geq 1 - \xi_i, \ \xi_i \geq 0 \ \forall i \tag{11}$$

where y_i is the class label value, which can be either $+1$ or -1, and x_i is the i th sample., will indicate the quantity of samples used. This problem dual form is used to solve it computationally,

$$\max \sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \ \forall \ i, \ \sum_i y_i \alpha_i = 0 \tag{12}$$

where $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$ is the kernel function that carries out the mapping that is not linear. Some of the most popular kernel functions are the Gaussian and Polynomial kernels.

3.2. Materials

3.2.1. Data set

In this work the Ethereum cryptocurrency has been selected for several reasons. According to some studies, it correlates with Bitcoin, so it is possible to use the latter as a benchmark. Indeed, ETH and BTC are the two biggest cryptocurrencies in terms of market capitalization, and they have led the cryptocurrency market for the last years. Ethereum's decentralized application development and e-contract capabilities have helped it become more well-known. In addition, other CCs can be significantly impacted by price changes and market trends, and researching Bitcoin and Ethereum can give important information about the sentiment of the market as a whole as well as potential price trends in the CC markets. They also have sizable developer communities, which adds to their dependability and stability [40]. Finally, interest in this cc has not stopped growing in recent years.

The historical Ethereum price data extracted from the Ethereum database was not only the closing price but also the high, low, open and adjusted close, although the one selected to be used in the prediction models has been the closing price of each day due to its relevance to calculate the metrics under study.

The time period goes from November 9, 2017, to May 13, 2022 (Fig. 1). The reason for choosing these dates has been the difficulty of finding data on some of the relevant characteristics of the evolution of ETH from previous dates. Bitcoin can be considered to have had a short lifespan so far, and Ethereum even shorter. The information from the initial moments, when these cryptocurrencies were launched on the market, adds a lot of noise. Therefore, to have reliable data we decided to use data from 2017. Furthermore, in view of the available information (Fig. 1), it can be seen that in the selected time period there are up to three different areas, one more stable and two others with a more abrupt evolution. This allows forecasting systems to be trained with very different behaviors.

The first period starts with a clear upward trend, multiplying its price by five. After reaching the maximum price at about €1500, towards the end of 2017, a downward trend begins with some ups and downs that continues until approximately December 2018. From then on, a second period of quasi-stability begins, where the price of Ethereum shows only very slight changes until April-May 2020. At this point the third period starts with a very pronounced price increase at the beginning of 2021. This is the time of highest volatility and greatest changes in the price, reaching what is until today the absolute historical maximum price of this cryptocurrency (11/13/2021), with more than €4600.

All data has been collected with daily granularity, having accumulated a total of 1647 datapoints per feature. Twenty-one features have been considered, of which two come from the sentiment analysis, although not all of them were eventually used after the feature selection procedure.

In addition, data preprocessing has been carried out to improve the training of the ML models, which has consisted of:

a) Fill in missing data: in general, values of all indicators or market indices are not available on weekends and non-working days in general. For those dates, the last known value has been used and the date has been inserted into the data structure.
b) Scale the data between 0 and 1, using the *MinMaxScaler* function from the *sklearn.preprocessing* library.

Due to the daily granularity of the data and its nature, neither the elimination of outliers nor any filtering has been necessary.

3.2.2. Information sources

With Python, the data collection procedure has been automated. The open-source library "yfinance" [41] has been utilized to retrieve historical data on cryptocurrencies and stock market indices. It provides a pythonic and threaded method of downloading market data from Yahoo! Finance. The data collected were obtained as follows,

- The Python module "requests" was used to collect the number of visits to the Wikipedia page for the word "Ethereum." It did this by sending an HTTP request straight to the Wikipedia API and receiving the data in an easy-to-parse JSON format that could be saved to our local disk as a *.csv file.
- The Python module "pytrends" has been utilized to retrieve historical Google statistics for the word "Ethereum." With the help of this library, we can obtain daily and monthly data that we may utilize to compile our historical data record through a straightforward interface to the Google backend.
- In order to perform sentiment analysis on daily news headers, the python libraries "*urllib*", "*requests*" and "*bs4*" were combined to first generate a local database, extracting strings that correspond to the

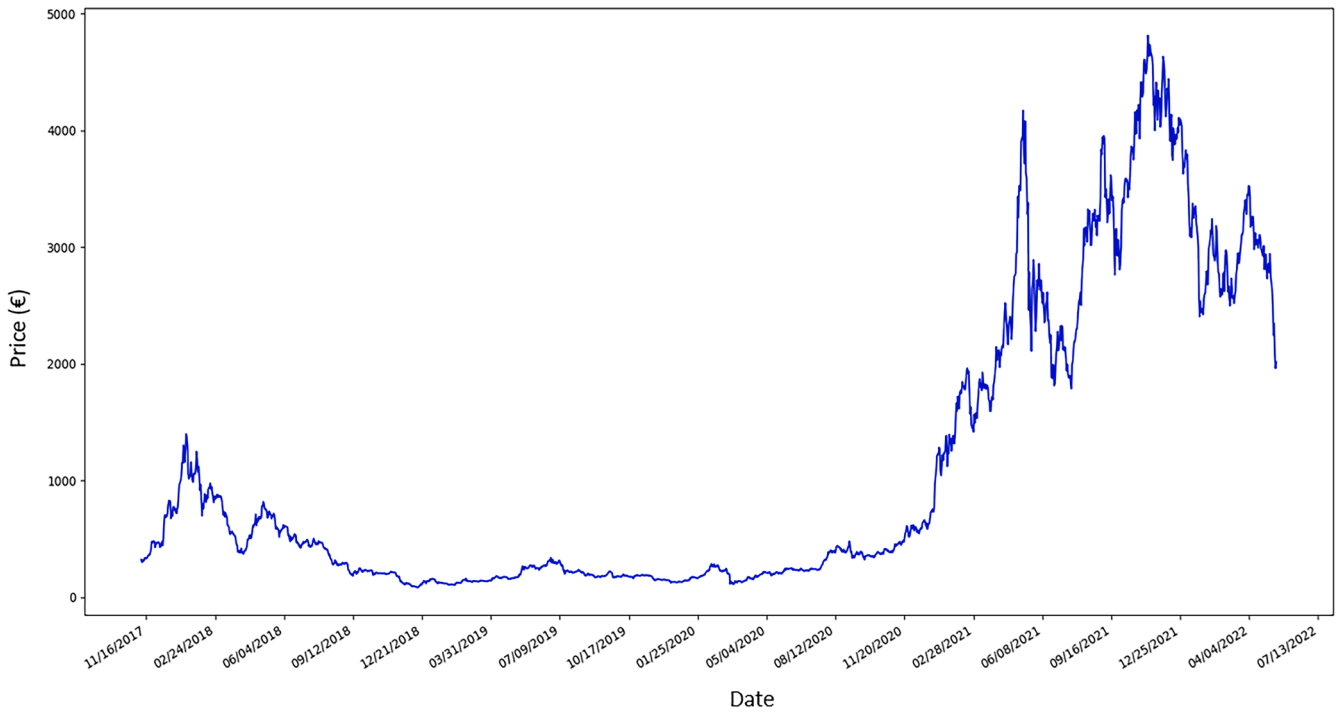


Fig. 1. Ethereum cryptocurrency close price evolution from 2017 to 2022.

mentioned news headers with daily granularity through Google news searches URLs.
- It was chosen to extract Reddit comments that contain the word "Ethereum" in order to do sentiment analysis. To do this, the "psaw" Python library was implemented. It is a simple wrapper for using the *pushshift.io* API to search public Reddit comments and submissions.
- EthereumScan: Some platforms do not provide the Ethereum Gas price and Ethereum transaction volume as variables. Their past information was manually obtained from the EtherScan site [42].

3.2.3. *Market indicators and forecasting metrics*

The selection of the market indicators, indices and other inputs have been inspired by other papers such as [24], resulting:

- Bitcoin price (BTC). Historical data of Bitcoin price with daily granularity.
- Relation Bitcoin-Ethereum (BtcEth). Historical data of the relation between Bitcoin and Ethereum prices with daily granularity.
- Standard & Poor's 500 index (SP500). It is a capitalization-weighted index, also known as value-weighted or market cap weighted. The simplest capitalization weighted index can be thought of as a portfolio consisting of all available shares of the stocks in the index. Dividing the portfolio market value by a factor, usually called the divisor, does the necessary scaling [43].
- VIX, or volatility index. This indicator calculates the amount of implied volatility in the bid/ask quotations of SPX options for the S&P 500 indicator over the following thirty days. Thus, while realized (or actual) volatility assesses the variability of historical (or known) prices, the VIX Index is a measure of the future [43].
- Commodity Bloomberg (BCOM) indicator. This index is intended to serve as a very liquid and well-rounded benchmark for investing in commodities. Since no single commodity or commodity sector dominates the Index, BCOM offers wide exposure to commodities as an asset class. When compared to non-diversified commodity baskets, the diversified commodity exposure of BCOM may lessen volatility [43].
- The DJI, or Dow Jones Industrial Average. Price-weighted index of thirty stocks that evaluates the performance of some of the biggest US corporations [43].
- The IXIC, or Nasdaq Composite. Almost all of the stocks listed on the Nasdaq stock exchange are included in the Nasdaq Composite, an index of the stock market. It is significantly biased in favor of businesses in the information technology industry [44].
- The MSCI World index from Morgan Stanley Capital International is a comprehensive global equity index that covers the performance of large and mid-cap stocks in each of the 23 developed market nations. It encompasses about 85% of each nation's market capitalization that has been adjusted for free float [45].
- HIS (Hang Seng Index). The biggest and most liquid stocks listed on the Main Board of the Stock Exchange in Hong Kong are included in the most generally quoted index of the Hong Kong stock market [46].
- The SSE is the Shanghai Stock Exchange Composite Index. It consists of all stocks listed on the SSE, including stocks denoted as A and B shares. In order to represent the price performance of listed stocks on the Shanghai Stock Exchange, the Index is weighted by total market capitalization [47].
- The N100, or Euronext 100 index. It is the Euronext NV pan-European exchange's blue chip index. It consists of the biggest and most actively traded stocks on Euronext. More than 20% of the issued shares of each stock must be traded during the rolling one-year analysis period. A size and liquidity examination of the investment universe is used to analyze the index on a quarterly basis [48].
- Relation Euro-US Dollar (EurUsd). Historical data of the relationship between Euro and USD prices with daily granularity.
- Crude Oil price (CrOil). This is one of the world's top trading commodities. It trades on the major commodities exchanges in the form of contracts. More details can be found in [49].
- Gold price (Gold). Commodity primarily traded in the New York Mercantile Exchange (NYMEX). It has the peculiarity that is not affected by consumption, meaning that most of the mined gold is still accessible. It has proven to be more stable, even though the difficult times [50].

- Daily visits to the English version of the Wikipedia entry “Ethereum” (WkViews).
- Google search daily volume for the term “Ethereum” (GTV). Number of searches in the U.S. of the term “Ethereum”, taken with daily granularity.
- Ethereum transactions fees price (EthGas). The unit used to quantify the computing effort (cost) needed to carry out particular tasks on the Ethereum network is called "gas." Every transaction has to be paid for.
- Ethereum daily transactions volume (ETV). Record of Ethereum transactions registered per day.

In addition, the quantitative metrics that will be applied to evaluate the forecast results are three values of the error, MAE (Mean Absolute Error), MSE (Mean Squared Error) and RMSE (Root-Mean Squared Error), as defined by Eq. (1).

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|e_i| MSE = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i)^2} \tag{13}$$

3.2.4. Feature selection

As an initial set of features, features that come from four sources have been selected. On the one hand, some from traditional markets (because they are correlated), others from the online world (google trends), sentiment analysis, and another group from the crypto world (Bitcoin, gas price, …). Seeing the information that was available and those that are most frequently used in similar works, 21 features have initially been considered. The selection of market indicators, indices and other features has been greatly influenced by other works [24], adding personal analysis and expert knowledge.

To reduce the dimensionality of the feature space, improve the prediction accuracy and reduce the computational cost, feature selection methods (FSM) should be applied [51]. But selecting the best feature selection method is not a simple task [13]. In this work, two FSMs have been applied for the regression approach and other two for classification.

For prediction using regression with NN these two FSM were used:

a. *Recursive Feature Elimination (RFE) with Random Forest Regressor*

Recursive Feature Elimination (RFE) functions as an ensemble classifier by combining several decision trees. The boot-strap resampling technique is used to randomly choose sample subsets for each decision tree throughout the training phase. The scores obtained from each decision tree define the ultimate categorization outcome of the RF. The ability of each tree to be classified as well as the correlation between the trees determine the classification error. The importance measure, which shows the influence of each predictor variable, is a crucial benefit of RF [52].

b. *Pearson’s Correlation Coefficient*

The well-known Pearson’s correlation coefficient is calculated by the product-difference method, i.e., first the means of the two variables is calculated separately, and then the deviations between the two variables and their respective means are also obtained, multiplying the two deviations to obtain the correlation coefficients, which reflect the degree of correlation between the variables.

For the classification with SVM, the feature selection techniques are:

c. *RFE-Random Forest Classifier*

The widely applied Random Forest Classifier is like the Random Forest Regressor but it uses the majority vote for the aggregation step.

d. *ANOVA*

One well-known statistical method for comparing many independent means is Analysis of Variance (ANOVA). By computing the ratio of variances across and between groups, the ANOVA method ranks features. The ratio shows the degree to which a feature and the group variables are related.

The selected features were obtained using the Python library *Scikit* for each FSM for five different prediction cases: one day, seven days, fifteen days, thirty days, and ninety days. As an example, Fig. 2 shows the ranking obtained by the RFE-RF regressor when applied to the market indicators for all those prediction horizons. As it is possible to see, some features are not useful in this application, for example "DJP" or "HSI". Others like N100 or VIX are clear candidates to be left out if you are looking for a group of characteristics that can work for all prediction periods. Some features like ’ETH’ or ’BTC’ seem to be possible inputs for most cases.

In summary, by applying FSM the number of model inputs can be reduced by approximately 50%. Furthermore, depending on the FSM and the prediction horizon, the subset of selected features changes. For both classification and regression approaches, the RFE method finds subsets with greater differences from each other. In three of the four FSMs analyzed it seems feasible to find a subset of characteristics that could be applied in all cases.

3.2.5. Sentiment analysis

The two features related to sentiment analysis have not been included in the feature selection process, since one of the goals of this study was to compare the performance of the intelligent systems with and without sentiment analysis. This extra characteristic was included in the forecasting. The sentiment analysis was meant to be extracted from Twitter, Reddit and Google news. After analysing the accessibility those platforms, Reddit and Google news were selected as we were able to automate the extraction of all their content and then, to parse it to a structured language such as XML, and work with it. Twitter was discarded due to its limitation to work with the professional APIs of the python packages available.

The queries were done in English language. There were comments that show clearly positive feeling/sentiment, and some of them more neutral or even negative headlines regarding the interests of Ethereum. To analyze each comment and identify a negative, positive or neutral sentiment, VADER (Valence Aware Dictionary and sentiment Reasoner) Sentiment Analysis was used. It is a lexicon-and rule-based sentiment analysis tool that performs well on texts across several domains and is particularly adapted to sentiments expressed on social media.

As a result of processing a comment, VADER returns four values: pos, neu, neg, and compound. The latter is the total of each word’s valence scores in the lexicon, standardized between $+1$ (the most extreme positive) and -1 (the most extreme negative) after being adjusted per the criteria. If you are looking for a single, unidimensional way to gauge the sentiment of a particular comment, this is the most helpful metric. It is accurate to refer to it as a "weighted and normalized composite score." The text passages that fit within each category are indicated by the pos, neu, and neg scores, which should all sum up to 1. In this work, when including the sentiment analysis, the compound value has been used as feature input of the models.

4. Price prediction. Regression approach with LSTM and GRU neural networks

Following the results obtained with the featured selection method, the prediction horizon considered is one day, seven days and fifteen days. Each set of features selected by the two FSM methods have been used as inputs to train each network, LSTM and GRU.

Tests have been carried out to find the best networks configuration. Finally, the network design is composed of a dense layer that reduces the

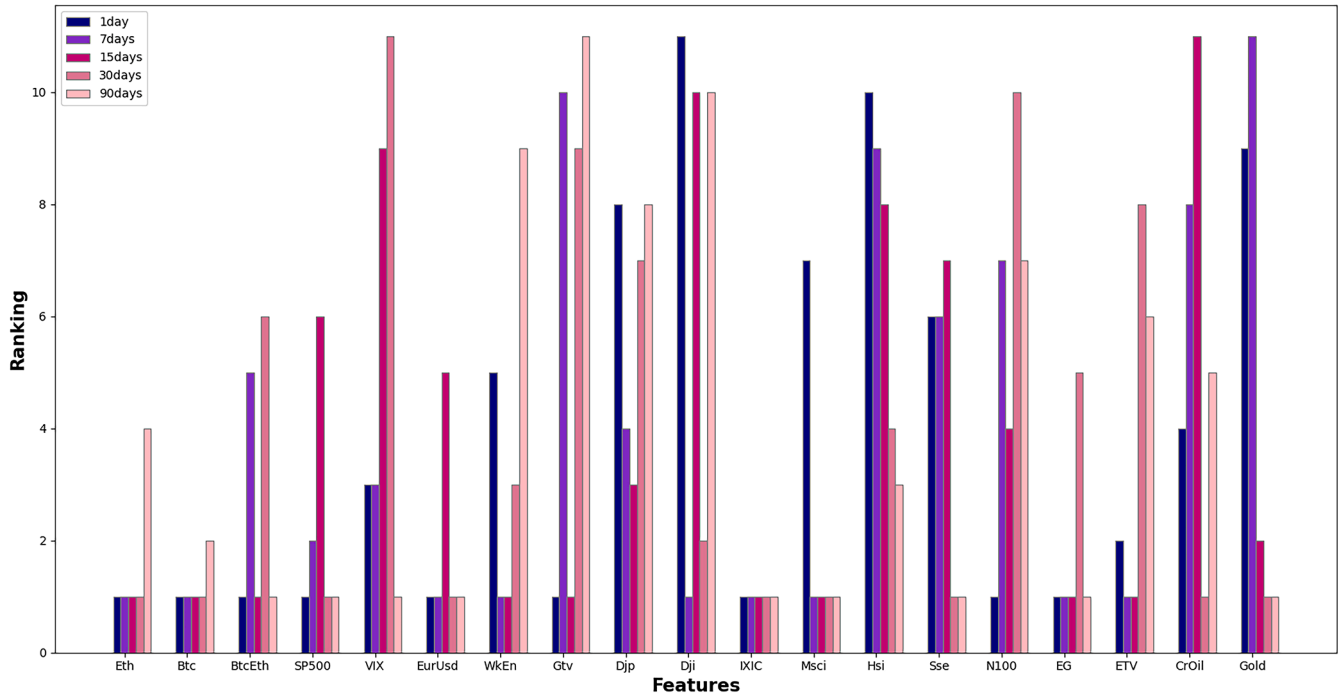


Fig. 2. Feature selection results with Recursive Feature Elimination (RFE) with Random Forest.

output to a single value, an input layer, a hidden layer, and dropout regularization with a rate of 0.2 that helps prevent overfitting. Both layers have 128 nodes. Depending on the model that is used, either GRU or LSTM will be used for the input and hidden layers. With a batch size of 32 and 60 epochs, the LSTM and GRU models were fitted using the mean squared error loss function and the effective Adam optimization.

Typically, it is advisable to use 70–80% of the data for training and the remaining 20–30% for testing. In this case, since the data is so limited, 20% of data for testing would mean having less than a year of time to validate the prediction, and using 30% would only consider the long period of greater stability in the price of Ethereum (Fig. 2) that leaves out the most significant changes. This has a negative impact on the network's ability to generalize. Therefore, it has been selected a shorter period of time, enough to provide the greatest amount of information possible and that still represents a good variety of the evolution of the price of Ethereum. Thus, due to the small amount of data and the way the variable to be predicted has evolved (Fig. 3, blue line), 650 samples out of 1647 have been chosen for training, or 39% of the total data. The training period is from July 19, 2021, until August 10, 2019. This time frame shows a good range of behavior in Ethereum's price evolution.

A collection of test data with two radically distinct periods remains

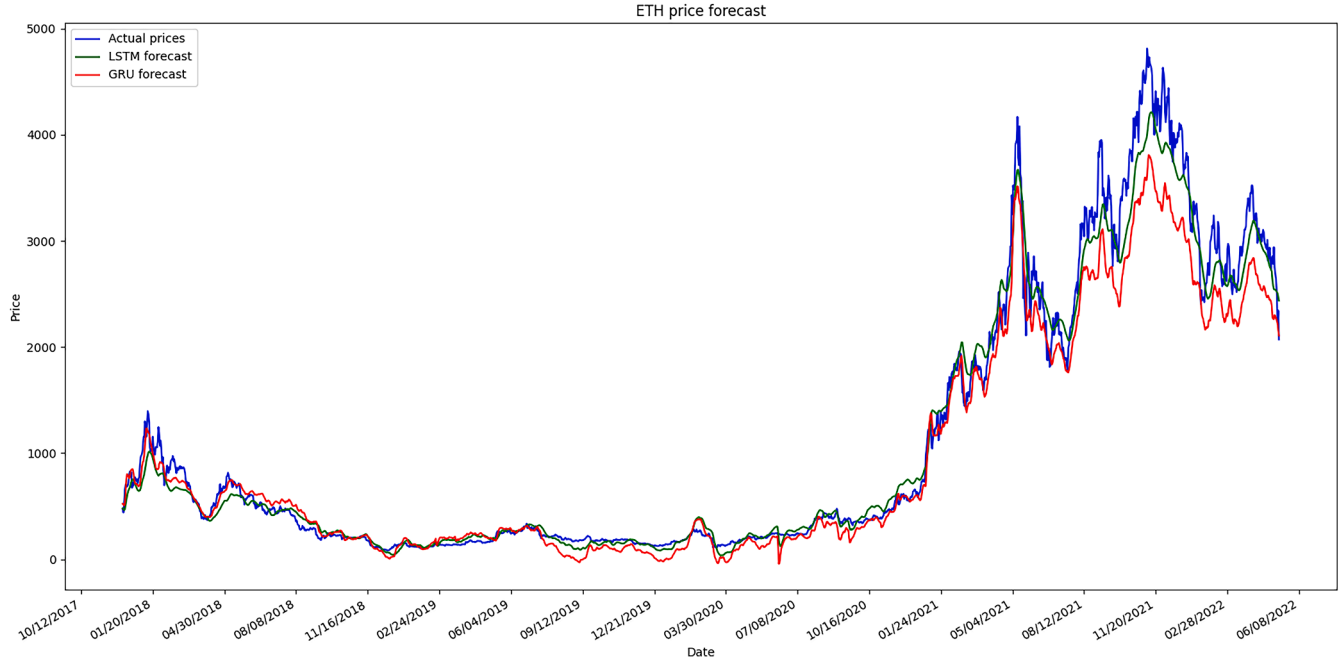


Fig. 3. ETH price prediction obtained with LSTM (green) and GRU (red) neural networks, and real values (blue).

after this selection of data for training (Fig. 3). In the first, there are over two years' worth of data with primarily minor fluctuations, and in the second, there are roughly ten months' worth of significant fluctuations in Ethereum's price. These two periods will be used and analyzed independently when applying the investment strategy.

In the simulations, 100 iterations are carried out and the prediction is the average of all of them.

Fig. 3 shows how the LSTM (green) and GRU (red) networks predict the price evolution of Ethereum (blue line, real values). Note that the time period from 08/10/2019 to 07/19/2021 has been used for training.

As said, the prediction of the ETH price evolution has been obtained for 1 day, 7 days, and 15 days, with and without sentiment analysis, considering as inputs of the LSTM and GRU networks the features selected by the RFE and the Pearson's correlation methods. After analysing in detail all the results obtained, the following conclusions can be drawn.

Case 1: 1-day forecasting:

- LSTM outperforms GRU in the first period, but GRU performs better in the second period.
- Features selected by the RFE method have better average performance. Only in the case of using the GRU network with sentiment analysis, the characteristics selected by the Pearson's method give better results for the second period.
- Predictions made with sentiment analysis are worse than without incorporating this feature.
- With or without sentiment analysis, both networks have minor errors in the first period, as expected by observing the evolution of prices in that time interval.

Case 2: Predictions for 7 days:

- GRU presents better results than LSTM when both use features selected by the RFE method, while LSTM predicts better than GRU with the features selected by Pearson's. This occurs for both periods and regardless of whether sentiment analysis is used.
- Predictions including sentiment analysis generally offer worse results.
- The results are worse than the predictions obtained for 1-day.
- In the second period, worse results are obtained than in the first, with and without sentiment analysis.

Case 3: 15-day predictions:

- GRU model with the characteristics selected by RFE technique obtains the best performance, both for the first and for the second period when sentiment analysis is not used.
- With sentiment analysis, LSTM with Pearson selected features is the best choice for the first period, but again GRU with RFE outperforms in the second period.
- Once again, predictions with sentiment analysis generally perform slightly worse than without this information.
- In line with the previous results, the second period shows greater deviations between the predicted and the actual prices.

In conclusion, in general it can be observed that:

- Features selected with RFE give lower prediction errors.
- With the RFE characteristics, the GRU network has the best performance for the three prediction horizon cases.
- No improvement has been observed using sentiment analysis; the errors have been even greater when including it.
- The results for the 1-day prediction are the best, while the 7-day and 15-day predictions show similar errors.
- The forecasts obtained for the first period have smaller errors than for the second period.

These conclusions summarize the quantitative results obtained with the predictions made for all possible cases. As an example, Table 1 shows the prediction obtained with GRU network for 1-day forecasting.

5. Price trend prediction. classification approach with svm

To predict the trend in the evolution of the price of ETH, a classification approach with the SVM technique is used. To do this, the price of this cryptocurrency has been subtracted from the price of Ethereum in each iteration by the number of subsequent days that we want to predict, that is, 1, 7 or 15. If the result of this difference is positive, it means that the price will be higher, and if it is negative, the opposite. This result has been encoded in Boolean variables as True and False, respectively, and has been provided as the expected result for training, as well as the expected result taken as a reference for the prediction evaluation.

The confusion matrix, which counts the true positive, false positive, true negative, and false positive items, and the F-measure, which is defined as a harmonic mean of precision (P) and recall (R), specified in (30), are the metrics used to assess the performance of the support vector machine.

$$F = \frac{2PR}{P + R} \tag{14}$$

The SVM has been configured with a linear kernel. It has been used to forecast with 1-day, 7day and 15-day horizons, with and without sentiment analysis, and using in all cases the features selected by the RFE and the ANOVA methods.

The results are shown in Table 2. The columns with and without mean with sentiment analysis or without sentiment data. The confusion matrix shows the in the first row the TN and FP, and in the second row the FN and TP, respectively.

Analyzing the results it can be seen that, in general, the characteristics selected with ANOVA allow us to obtain slightly better precision. Furthermore, the use of sentiment analysis data does not provide any improvement, nor a clear worsening of the predictions. Furthermore, there is no clear conclusion about the influence of the prediction horizon. The results are generally very similar and around 50%, so it cannot be said that this is a good approach for these predictions.

6. Knowledge-based investment strategies

One important goal is to have a solid investment strategy that supports the data forecasting and help making decisions based on the knowledge extracted. Thus, it is verified that utilizing machine learning techniques can contribute to an effective investing system. An investment strategy has been developed for each of both, the regression and classification methods. The metrics used to study the performance of investment strategies, common to both approaches, are the following.

Table 1
Error metrics (MSE, RMSE, MAE) for 1-day price prediction without sentiment analysis using variables from RFE and Pearson's feature selection technique, with LSTM and GRU neural networks.

	LSTM			GRU		
	MSE	RMSE	MAE	MSE	RMSE	MAE
1st period						
RFE	0.045	0.207	0.015	0.057	0.228	0.019
Pearsonś	0.095	0.298	0.024	0.157	0.379	0.033
2nd period						
RFE	1.382	1.152	0.101	1.152	1.050	0.092
Pearsonś	1.688	1.247	0.111	1.236	1.081	0.097

Table 2
F1-score and confusion matrix for 1, 7 and 15 days ETH trend prediction with SVM, with and without sentiment analysis.

	RFE-RFC				ANOVA			
	without		with		without		with	
	f1-score: 0.5		f1-score: 0.51		f1-score: 0.48		f1-score: 0.51	
	Confusion matrix		Confusion matrix		Confusion matrix		Confusion matrix	
1 day	604	230	563	271	512	322	581	253
	583	198	516	265	516	256	543	238
7 days	without		with		without		with	
	f1-score: 0.49		f1-score: 0.48		f1-score: 0.51		f1-score: 0.51	
	Confusion matrix		Confusion matrix		Confusion matrix		Confusion matrix	
	478	368	544	302	290	556	413	433
	451	312	540	223	235	528	351	412
15 days	without		with		without		with	
	f1-score: 0.5		f1-score: 0.47		f1-score: 0.52		f1-score: 0.49	
	Confusion matrix		Confusion matrix		Confusion matrix		Confusion matrix	
	485	354	509	330	318	521	375	464
	440	322	517	245	251	511	359	403

- Earnings: It is the difference between the final and the initial capital. It is positive in case of obtaining profit or negative is there is a loss. That is,

$$\text{Earnings} = \text{Initialfiat} - -\text{Finalfiat} \tag{15}$$

- % Earnings, expressed in%, means the percentage over the initial fiat that are earnings. That is,

$$\%\text{Earnings} = (\text{Earnings} * 100)/\text{Initialfiat} \tag{16}$$

- Max drawdown: maximum difference between the maximum amount of capital owned and the minimum amount of capital earned after that maximum and before the capital surpasses the mentioned maximum. Given by equation (26) where P = Local maximum value, L = Local minimum value reached before the previous P is surpassed, and i = iteration of all local maximum values in the series.

$$\text{Maxdrawdown} = \text{Max}[(\text{P}_\text{i} - \text{L}_\text{i}) / \text{P}_\text{i}] \tag{17}$$

- Profit factor *Pf: Difference between the accumulated profit and the accumulated loss. It relates the amount of profit per unit of risk, with values greater than 1 indicating a profitable system, where P means the profit when an operation has been profitable; L: loss when an operation has ended with loss.

$$\text{Pf}^* = \frac{\sum_{i=1}^{n} P_i}{\sum_{k=1}^{m} L_k} \tag{18}$$

- Average profit: Profit average per operation, considering as an operation the process of buying Ethereum and selling it back.

$$\text{AvgP} = \frac{\sum_{i=1}^{n} P_i}{nOperations} \tag{19}$$

- Average loss (per operation), considering as an operation the process of buying Ethereum and selling it back.

$$\text{AvgL} = \frac{\sum_{k=1}^{m} L_k}{nOperations} \tag{20}$$

In addition, the number of buying operations, selling operations, stoploss triggered and final fiat will be obtained for each case.

6.1. Investment strategy based on the ETH price forecasting with the regression approach with GRU y LSTM neural networks

This investment strategy uses the ETH price prediction obtained with the regression approach (Fig. 3). By studying the peaks two days before and three days after, the method finds local maxima or minima in the prediction data. The search will be performed for each iteration (t) on the third day before the current iteration (t-3). Following this logic, the third day after the maximum or minimum found will be, precisely, the day of the current iteration. To accept the maximum or minimum in the forecast data, the day of the current iteration must also follow the trend of the previous 2 days, to confirm that the peak found is indeed stable (Fig. 4). If a high is identified, a sell order is placed on the prediction horizon date, in the case of Fig. 4, 7 days. Likewise, if a minimum is identified, a purchase order is placed for 7 days later.

To do this, the system uses the forecast data received from the previous step and will iterate them looking only at the past forecast data and the "today" data, without ever making use of the forecast data for future days since in a real scenario these data would not be available. Today's forecast data is understood to be the value that the price of Ethereum will have on the future day that has been defined (it can be 1 day, 7 days or 15 days) and, therefore, the peak found three days before the current iteration day also corresponds to the future. Of course, this does not apply to forecasting for one day in the future. In that specific case, the system places the order as soon as possible after the maximum or minimum has been recognized, which is the day after the current iteration day. This search for maximums and minimums begins on the fifth day of available data (Fig. 4).

In addition, a stoploss strategy has been implemented and, after some initial tests, it was set to 7% of the value that Ethereum had at that time. To select this value, the balance between two factors was taken into account: first, the number of stoploss that are activated when executing the investment strategy, trying to reduce them as much as possible so that the investment strategy selects the moments in which to sell; and second, avoid large losses when the investment strategy could not correctly predict a price drop.

The stoploss increases every time the value of Ethereum increases. The value of Ethereum is considered in this strategy as virtual fiat and is recalculated every day according to its new price. If the new virtual fiat is less than the previously calculated stoploss, a selling order is activated that same day, transferring all the virtual fiat to real fiat, since all Ethereum is sold.

In this strategy there will always be either only fiat or only Ethereum (virtual fiat), without having any progressive buying system based on trust or a selling system that guarantees a minimum profit.

This procedure is shown in Fig. 4. Points 1 and 4 are a maximum and a minimum but they would not be considered in this strategy since three iterations after, the price of Ethereum changes its trend. However, points 3 and 5 are maximums and 2 and 6 are minimums, all of them fulfilling the conditions indicated for placing selling and buying orders, respectively.

In the horizontal axis it has been represented the delays that will be repeated over time, finding the peaks and placing the orders. In this case the minimum is found at 6, which corresponds to t-3; with a prediction 7 days later the buying order will be placed at $t + 4$. In this way, four days later the strategy will invest all the fiat money it has in the buying of Ethereum.

This inversion strategy has been applied to the predictions obtained with the two networks, LSTM and GRU, for the different prediction horizons, and with and without sentiment analysis. It has been run 100 times, and the average results have been calculated. Table 3 shows an example.

The conclusions that can be drawn from all the cases simulated are

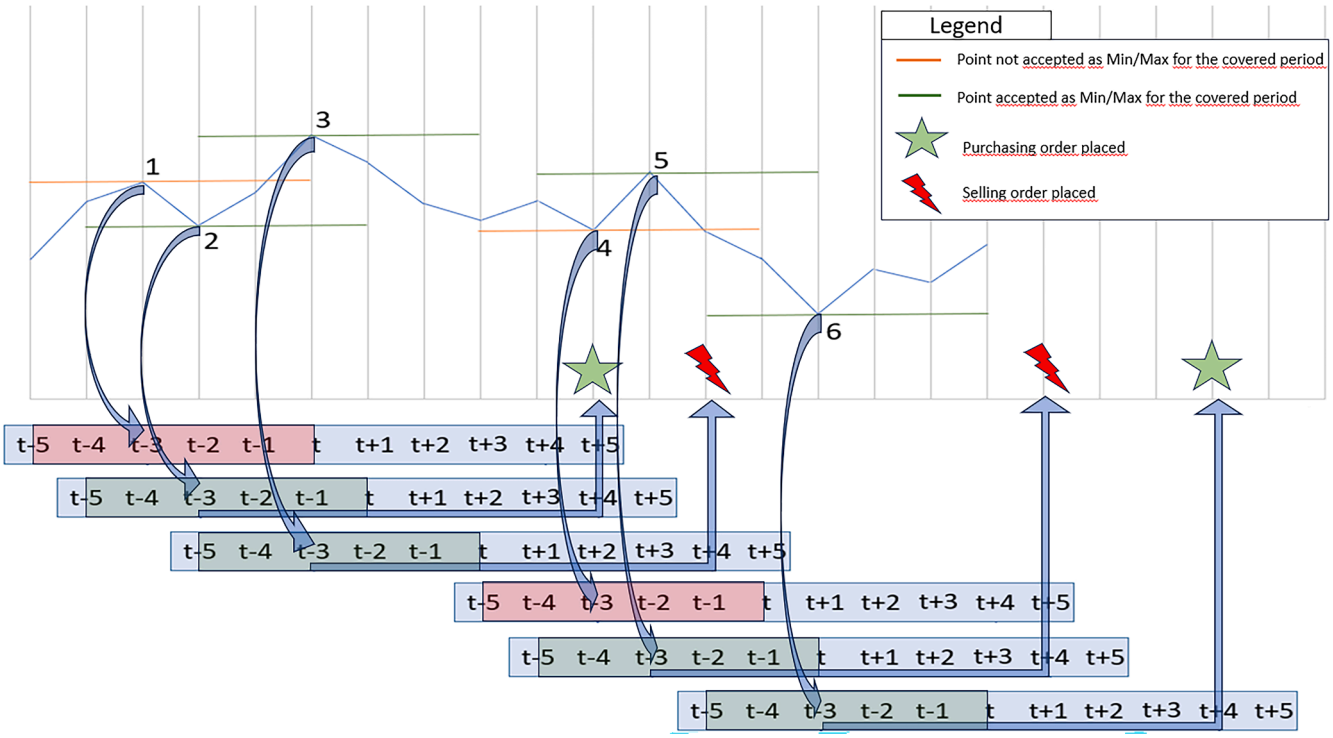


Fig. 4. Graphical representation of the regression investment strategy based on ETH price prediction. Time delays are represented in the horizontal axis. Maximums and minimums found are represented by green lines.

Table 3
Results of the regression investment strategy using 1-day LSTM and GRU neural networks ETH price prediction with Pearson's features, with sentiment analysis.

Pearson's	1st period		2nd period	
Network	LSTM	GRU	LSTM	GRU
Final Fiat	28,134.64	47,297.83	14,813.86	10,696.88
Earnings	18,134.64	37,297.83	4813.86	696.88
% Earnings	181.34	372.97	48.13	6.96
Max. Drawdown (%)	31.94	38.27	16.98	28.45
Profit factor	2.10	2.15	5.16	1.14
Avg. Profit	948,6	1290,2	809,56	389,7
Avg. Loss	460,62	594,36	188,94	344,98
Buy Operations	38.83	54.81	7.83	16.68
Sell Operations	38.82	54.65	7.83	16.68
StopLoss	18.73	18.85	4.16	7.27

the following.

Case 1: 1-day forecasting:

In most cases the strategy has been highly profitable. The profit obtained by using the GRU network combined with the characteristics selected by the Pearson's method during the first period is surprising and very remarkable. It is also worth mentioning that the GRU network has generated a higher trading volume compared to LSTM, and that the LSTM network seems to be more suitable for the second period, while GRU performs better in the first period. That is, the LSTM network performs better in conditions of greater volatility while GRU offers better performance in the periods, when the price of Ethereum presents a more stable evolution. Although in general the results are very positive, with great profits, there was one case that returned losses (GRU with RFE in the second period).

Looking for the best overall performance, the best option would be to use the features selected by Pearson's with an LSTM network, since it has a very solid performance in terms of risk, as can be seen through the calculated profit factor. On the other hand, if greater profits are sought but with greater risk, the GRU network together with the Pearsons features would be the indicated option.

In the metrics evaluated, including sentiment analysis does not improve the results.

Case 2: 7-day forecasting:

With the 7-day forecast although the results are very positive, lower profits are achieved in the first period than with the 1-day forecast, while for the second period they are on average higher with the GRU network, since with LSTM they are very similar to the previous ones.

The profit factor worsens in the first period with both networks, while it is clearly better in the second period when using GRU, with no cases of loss.

GRU combined with Pearson's again returns the best earnings, this time with slightly higher risk than LSTM.

Using sentiment analysis is clearly not an improvement over not using it. However, it does improve the earnings obtained by LSTM with Pearson's characteristics and becomes the best option for any 7-day forecast.

The most notable improvement is the excellent profit factor obtained for the second period when using LSTM with Pearson's, which shows an extremely strong and reliable performance of the strategy, although for the first period it could be defined almost as weak.

Case 3: 15-day forecasting:

For 15-day predictions, the results are the worst. However, in the second period the earnings are as good as before or even better. It seems that profits increase when using long-term predictions during times of higher volatility, while in the short term, when the price barely changes, the opposite is true.

Once again, combining profit factor and earnings metrics, the best option is LSTM with Pearson's with a conservative approach, and GRU if

the objective is to obtain more profits at the cost of greater risk.

Although some specific cases show an improvement, the application of sentiment analysis generally gives worse results.

In general, both LSTM and GRU can offer very good performance but with some clear differences. Speaking of Max Drawdown, this metric reveals that both strategies could be improved for the first period, where the LSTM ends up with an average of ~32% for 15-day and 7-day predictions, and 26% for 1-day predictions, and GRU with an average of 39%, 51% and 33% for 15-day, 7-day and 1-day predictions, respectively. These results cannot be considered good and it would be necessary to study how to avoid losing capital once earned. Perhaps modifying the stop-loss percentage could be a good starting point.

Almost always the use of sentiment analysis impoverishes the performance of both models, although in the second period, with greater volatility, the results are much better and could be considered acceptable, obtaining an average of less than 15% with LSTM and less than 20% with GRU for the Max Drawdown. It is also interesting to mention that the average profit per operation is better for short-term predictions where the evolution of the ETH price is stable and, as the prediction horizon increases, the best performance moves towards periods of greater volatility (2nd period in this case).

Finally, it can be concluded that the best combination would be an LSTM model with features selected by the Pearson's method without sentiment analysis data.

6.2. *Investment strategy based on the ETH trend forecasting with the SVM classification approach*

Using the data from the ETH price trend prediction (obtained with SVM), a new simple investment strategy has been designed. The strategy consists of placing a buy order when the price trend is positive. Once the purchase of Ethereum is made, the strategy blocks new buying or selling orders for the forecast interval used, for example, 7 days. At the end of that period of time, the future trend will be analyzed again and if the estimate is that the price rises again, Ethereum is not sold again to fiat; otherwise, if the trend is for the price to decline, Ethereum is converted back to fiat, again blocking potential orders to buy Ethereum during the forecast period, and so on.

Additionally, a stoploss technique has been implemented so, even during the blocked time period, if the price drops by a certain percentage, Ethereum will be converted back to fiat. And if the price of Ethereum increases, the stoploss is updated with a new higher value.

In Fig. 5 it is possible to see how the first iteration has a "true" value, that is, the prediction estimates that the price of ETH will be higher in the next seven days, so the order would be to invest in Ethereum. The next estimate is made seven days later, with a "false" forecast result, indicating that the price of Ethereum will be lower the next seven days; therefore, the Ethereum would be sold back to fiat money in this iteration and the second lock-up period begins. If during the first blocked period, where due to the first "true" value all our assets are Ethereum, the value of the virtual fiat drops below the stoploss, the Ethereum would immediately be sold back to fiat, without waiting for the end of the blocked period.

As was done with the regression approach, this investment strategy has been evaluated to see if it is possible to obtain profits in the two time periods already presented above with the three prediction horizons proposed. The results are summarized below.

Case 1: 1-day forecasting:

Like the investment system with regression, the first period has more trades than the second period with RFE feature selection. With ANOVA, the strategy is very different since it has a single operation. In any case, this second scenario is the most convenient, analyzing the profits obtained in both periods and calculating the final fiat that would result from the combination of both. But overall it is not a good way to invest.

Using the prediction that includes sentiment analysis the strategy is profitable using the features selected with RFE. The profits in both periods exceed 20%, which is a good result. However, with the FSM ANOVA the prediction does not work well with this strategy, and although the second period yields a very good profit, more than 20% of the initial investment is lost in the first period. This would lead to ruling out this option.

Case 2: 7-day forecasting:

With ETH trend prediction for 7 days without sentiment analysis, this strategy performs quite poorly, giving losses of approximately 20–50%, except in one case that is profitable. Therefore, it is not an interesting option.

Using sentiment analysis data generates very different behavior for this strategy. It triggers many trades with ANOVA FSM and this generates high profits with the help of stoploss mechanism. However, in the second period that shows greater volatility, depending on the time horizon in which this strategy is applied, there will be a significant risk of losing part of the investment. Therefore, although sentiment analysis once again helps to improve the performance of this strategy, it would not be advisable to use it.

Case 3: 15-day forecasting:

The investment strategy with the 15-day prediction of the Ethereum price trend does not make profits in any case. As with 1-day and 7-day predictions, and unlike what was observed with regression, sentiment analysis helps improve strategy performance. In this case again the ANOVA method together with the sentiment data improves the results. In any case, although using this strategy could be profitable, it would be very risky due to the low performance it offers in situations such as during the second period of high volatility, where there is a negative result of more than 15%.

Table 4 shows an example of the results obtained with this investment strategy.

As a summary, predicting ETH price trends is not beneficial in general. As the prediction results announced, the implemented investment strategy obtains profits in about 50% of the cases and losses in the other half. The two FSMs give results with too high risk. Again, the addition of sentiment analysis has not added any positive value to the final results.

Regarding the prediction results, there is some consensus that due to its randomness, no method is robust enough to predict the price of cryptocurrencies [53]. In this paper the authors apply CNN-LSTM price

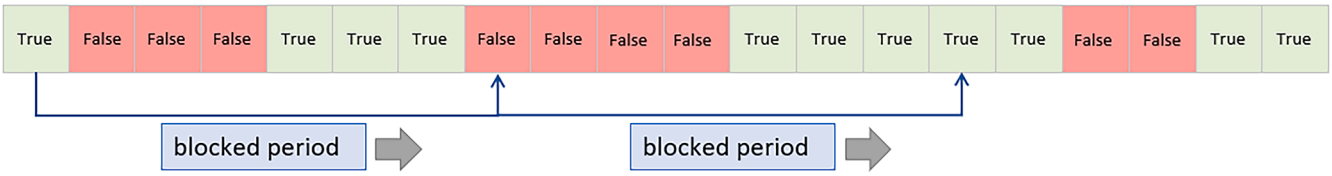


Fig. 5. Graphical representation of the classification investment strategy based on ETH trend prediction. True means the predicted value of the ETH will be higher thus the order is "buy". False represents a lower estimated price of the ETH, the order is "sell".

Table 4
Results of the classification investment strategy using 1-day ETH trend prediction with ANOVA features, with sentiment analysis.

ANOVA Network	1st period SVM	2nd period SVM
Final Fiat	7805.64	16,346.61
Earnings	−2194.35	6346.61
% Earnings	−21.94	63.46
Max. Drawdown (%)	46.03	22.21
Profit factor	0.83	1.97
Avg. Profit	185.08	754.44
Avg. Loss	220.76	381.12
Buy Operations	62	17
Sell Operations	61	17
StopLoss	0	17

prediction, including Twitter volume and Twitter sentiment, to predict the close price of BTC and ETH using data from 2013 to 2017 to 2019, respectively. They value this empirical analysis as a good basis for cryptocurrency investment but do not propose any investment strategy.

These results have been reached after considering a series of parameters that can influence the final outcomes. In fact, different values of some of the parameters of the proposed systems have been tested. Specifically, this sensitivity analysis has covered:

- The prediction horizon of the forecasting systems. Predictions of 1, 7, 15, 30 and 90 days have been analyzed. The long term has been ruled out given that cryptocurrency prices in general, and the price of Ethereum in particular, have proven to be very sensitive to market changes and daily events in the global economy and geopolitics.
- Sentiment analysis. The influence of including this feature has been studied in the various proposed scenarios.
- The structure of the data set for training and testing.
- Capital owned in traditional currency (USD) and in cryptocurrency, specifically ETH.
- The different characteristics that can be used to train the systems.

This has allowed us to obtain general systems, with a methodology that can be applied to other markets and other currencies.

7. Conclusions and future works

The main contribution of this work is the study and a better understanding of the behaviour of the cryptocurrency market and the factors that may influence it. Every advance in knowledge of this market, which is still young, is relevant and has important implications for the countries' economies since it is expanding and gaining value continuously in recent years. In addition, as it grows, an increasingly close connection is being formed with conventional stock markets, opening the door to new types of investment systems. Market analysis and the value of prediction have been highlighted.

Another objective of this study was to test the validity of the application of ML techniques to data that comes from this highly volatile market. It has been shown how neural networks (specifically, LSTM and GRU) and SVM allow us to forecast the price and trend of a cryptocurrency, Ethereum, with sufficient precision to be able to use these predictions as information for financial investment.

This has been proven by designing two investment strategies that achieve profits using the results of the prediction models in a real scenario.

These findings lay the foundation for future developments. The influence of the feature selection process, as well as the prediction horizon, has been analysed. The relevance of selecting the appropriate features to obtain a good ETH price prediction has also been demonstrated, as well as the importance of training with a small but necessarily very varied amount of data due to the high volatility of this cryptocurrency.

The effect of including sentiment analysis in prediction has also been explored. In general, it has not been helpful in predicting the price of ETH, perhaps due to very simple techniques being followed to extract data from online platforms and the lack of analysis from relevant sources such as Twitter. This is one of the limitations of this study that has possible improvement. Other non-free sources could be used to improve sentiment analysis. The requirements of some tools to process information from news headlines have also been a limitation. On the other hand, something that cannot be lost sight of is the dependence of the news about this market on the country in which it is operating.

Some relevant implications of the findings obtained in this study focus on their possible application for the cryptocurrency market in general.

These trained networks and investment strategies can be used for most cryptocurrencies known as altcoins, as they are highly correlated and influenced mainly by the same factors, especially the evolution of BTC. This opens the door to predicting and investing in young cryptocurrencies that do not yet offer a large amount of historical data, which may be insufficient to train the NNs. The very fact of proposing an investment strategy is already an aid for decision making.

Furthermore, the designed investment tools can be computationally implemented and applied in the real world. They not only support decision making but are also useful for investigating the market, the importance of some features and analysing trends. They can be very useful for working with new currencies that may be launched, as well as for investing and making money.

Another line of future work that would improve some of the limitations of the proposed systems would be to obtain more historical data to upgrade the training of the ML techniques and achieve better prediction results. It is expected that as time goes by there will be more data that can be used. It is further anticipated that as the relationship between cryptocurrency markets and conventional stock markets grows, the volatility of cryptocurrency values will decrease and, as a result, the evolution of these prices will become more predictable.

The application of other deep learning techniques with larger amounts of data, combined with appropriate algorithms, could also be explored, which could help reduce risks and increase profits when investing in cryptocurrencies.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Adrián Viéitez: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Matilde Santos:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Rodrigo Naranjo:** Validation, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] F.J. García-Corral, J.A. Cordero-García, J. de Pablo-Valenciano, J. Uribe-Toril, A bibliometric review of cryptocurrencies: how have they grown? Financ. Innov. 8 (1) (2022) 1–31, https://doi.org/10.1186/s40854-021-00306-5.

[2] I.M. Sifat, A. Mohamad, M.S.B.M. Shariff, Lead-lag relationship between bitcoin and ethereum: evidence from hourly and daily data, Res. Int. Bus. Finance 50 (2019) 306–321.

[3] M. Kutera, Cryptocurrencies as a subject of financial fraud, J. Entrepreneur., Manag. Innov. 18 (4) (2022) 45–77.

[4] Buterin, V. (2013). Ethereum white paper: a next generation smart contract & decentralized application platform. https://github.com/ethereum/wiki/wiki/White-Paper, last accessed on 28/01/2024.

[5] B. Wu, Analysis of Ethereum ghost protocol under blockchain framework, Highlights Sci., Eng. Technol. 60 (2023) 121–127.

[6] V. Buterin, Proof of stake: The making of Ethereum and the Philosophy of Blockchains, Seven Stories Press, 2022.

[7] R. Naranjo, J. Arroyo, M. Santos, Fuzzy modeling of stock trading with fuzzy candlesticks, Expert. Syst. Appl. 93 (2018) 15–27.

[8] Binance (2024). Price evolution of Ethereum under the market view of Binance, available at: https://www.binance.com/en/price/ethereum, last accessed on 28/01/2024.

[9] R. Naranjo, M. Santos, A fuzzy decision system for money investment in stock markets based on fuzzy candlesticks pattern recognition, Expert. Syst. Appl. 133 (2019) 34–48.

[10] N.M. Farimani, P. Parsafar, S. Mohammadi, Evaluation performance of time series methods in demand forecasting: box-Jenkins vs artificial neural network (Case study: automotive Parts industry), J. Stat. Comput. Simul. 92 (17) (2022) 3639–3658.

[11] M. Hadwan, B.M. Al-Maqaleh, F.N. Al-Badani, R.U. Khan, M.A Al-Hagery, A hybrid neural network and box-jenkins models for time series forecasting, CMC-Comput. Mater. Contin 70 (2022) 4829–4845.

[12] S.K. Purohit, S. Panigrahi, Novel deterministic and probabilistic forecasting methods for crude oil price employing optimized deep learning, statistical and hybrid models, Inf. Sci. (Ny) 658 (2024) 120021.

[13] H.H. Htun, M. Biehl, N. Petkov, Survey of feature selection and extraction techniques for stock market prediction, Financ. Innov. 9 (1) (2023) 26.

[14] A. Kurani, P. Doshi, A. Vakharia, M. Shah, A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting, Ann. Data Sci. 10 (1) (2023) 183–208.

[15] D.W. Jorgenson, M.L. Weitzman, Y.X. ZXhang, Y.M. Haxo, Y.X. Mat, Can neural networks predict stock market? AC Invest. Res. J. 220 (44) (2023).

[16] N. Latif, J.D. Selvam, M. Kapse, V. Sharma, V. Mahajan, Comparative performance of LSTM and ARIMA for the short-term prediction of bitcoin prices, Austr. Account., Bus. Finance J. 17 (1) (2023) 256–276.

[17] D.F. Gerritsen, E. Bouri, E. Ramezanifar, D. Roubaud, The profitability of technical trading rules in the Bitcoin market, Financ. Res. Lett. 34 (2020) 101263.

[18] C. López-Martín, S. Benito Muela, R. Arguedas, Efficiency in cryptocurrency markets: new evidence, Eurasian Econ. Rev. 11 (3) (2021) 403–431.

[19] I. Nasirtafreshi, Forecasting cryptocurrency prices using recurrent neural network and long short-term memory, Data Knowl. Eng. 139 (2022) 102009.

[20] G. Dudek, P. Fiszeder, P. Kobus, W. Orzeszko, Forecasting cryptocurrencies volatility using statistical and machine learning methods: a comparative study, Appl. Soft. Comput. 151 (2024) 111132.

[21] E. García-Gonzalo, P.J. García Nieto, J. Garcia Rodríguez, F. Sánchez Lasheras, G. Fidalgo Valverde, A support vector regression model for time series forecasting of the COMEX copper spot price, Log. J. IGPL. 31 (4) (2023) 775–784, https://doi.org/10.1093/jigpal/jzac039.

[22] R. Gupta, J.E. Nalavade, Metaheuristic assisted hybrid classifier for bitcoin price prediction, Cybern. Syst. 54 (7) (2023) 1037–1061.

[23] S. Hansun, A. Wicaksana, A.Q. Khaliq, Multivariate cryptocurrency prediction: comparative analysis of three recurrent neural networks approaches, J. Big. Data 9 (1) (2022) 1–15.

[24] F. Sabry, W. Labda, A. Erbad, Q. Malluhi, Cryptocurrencies and artificial intelligence: challenges and opportunities, IEEe Access. 8 (2020) 175840–175858, https://doi.org/10.1109/ACCESS.2020.3025211.

[25] M.A. Ammer, T.H. Aldhyani, Deep learning algorithm to predict cryptocurrency fluctuation prices: increasing investment awareness, Electronics. (Basel) 11 (15) (2022) 2349.

[26] S. Chen, Cryptocurrency financial risk analysis based on deep machine learning, Complexity. 2022 (2022), https://doi.org/10.1155/2022/2611063. Article ID 2611063.

[27] A.H. Al-Nefaie, T.H. Aldhyani, Bitcoin price forecasting and trading: data analytics approaches, Electronics. (Basel) 11 (24) (2022) 4088.

[28] A. Dutta, S. Kumar, M. Basu, A gated recurrent unit approach to bitcoin price prediction, J. Risk. Financ. Manage 13 (2) (2020) 23, https://doi.org/10.3390/jrfm13020023.

[29] V. Derbentsev, V. Babenko, K.I. Khrustalev, H. Obruch, S.O. Khrustalova, Comparative performance of machine learning ensemble algorithms for forecasting cryptocurrency prices, Int. J. Eng. 34 (1) (2021) 140–148.

[30] W. Yiying, Z. Yeze, Cryptocurrency price analysis with artificial intelligence, in: 2019 5th International Conference on Information Management (ICIM), IEEE, 2019, pp. 97–101, https://doi.org/10.1109/INFOMAN.2019.8714700.

[31] J. Wang, F. Ma, E. Bouri, Y. Guo, Which factors drive Bitcoin volatility: macroeconomic, technical, or both? J. Forecast. 42 (4) (2023) 970–988.

[32] P. Jaquart, D. Dann, C. Weinhardt, Short-term bitcoin market prediction via machine learning, J. Finance Data Sci. 7 (2021) 45–66.

[33] S. Sharma, A. Majumdar, Deep state space model for predicting cryptocurrency price, Inf. Sci. (Ny) 618 (2022) 417–433.

[34] L. Felizardo, R. Oliveira, E. Del-Moral-Hernandez, F. Cozman, Comparative study of bitcoin price prediction using wavenets, recurrent neural networks and other machine learning methods, in: 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), IEEE, 2019, pp. 1–6, https://doi.org/10.1109/BESC48373.2019.8963009.

[35] F. Colon, C. Kim, H. Kim, W. Kim, The effect of political and economic uncertainty on the cryptocurrency market, Financ. Res. Lett. 39 (2021) 101621, https://doi.org/10.1016/j.frl.2020.101621.

[36] F. Valencia, A. Gómez-Espinosa, B. Valdés-Aguirre, Price movement prediction of cryptocurrencies using sentiment analysis and machine learning, Entropy 21 (6) (2019) 589, https://doi.org/10.3390/e21060589.

[37] E. Bouri, R. Gupta, Predicting Bitcoin returns: comparing the roles of newspaper- and internet search-based measures of uncertainty, Financ. Res. Lett. 38 (2021) 101398.

[38] Z. Wen, E. Bouri, Y. Xu, Y. Zhao, Intraday return predictability in the cryptocurrency markets: momentum, reversal, or both, North Am. J. Econ. Finance 62 (2022) 101733.

[39] O. Kraaijeveld, J. De Smedt, The predictive power of public Twitter sentiment for forecasting cryptocurrency prices, J. Int. Financ. Markets, Inst. Money 65 (2020) 101188, https://doi.org/10.1016/j.intfin.2020.101188.

[40] K. Mokni, G. El Montasser, A.N. Ajmi, E. Bouri, On the efficiency and its drivers in the cryptocurrency market: the case of Bitcoin and Ethereum, Financ. Innov. 10 (1) (2024) 39.

[41] Yfinance (2024) project website, https://pypi.org/project/yfinance/, last accessed on 28/01/2024.

[42] Etherscan (2024), available at https://etherscan.io/, last accessed on 28/01/2024.

[43] Cboe (2023), at: https://www.cboe.com/tradable_products/vix/faqs/, last accessed on 28/01/2024.

[44] Nasdaq (2024). English version of the Wikipedia entry "Nasdaq Composite", available under: https://en.wikipedia.org/wiki/Nasdaq_Composite, last accessed on 28/01/2024.

[45] MSCI (2020). MSCI Index Calculation Methodology (2020), available under: https://www.msci.com/eqb/methodology/meth_docs/MSCI_IndexCalcMethodology_Jan2020.pdf, last accessed on 28/01/2024.

[46] HSI (2023) Hang Seng Index Factsheet, available at www.hsi.com.hk/static/uploads/contents/en/dl_centre/factsheets/hsie.pdf, last accessed on 28/01/2024.

[47] SSE (2023) (Shangai stock exchange) indices calculation & maintenance, available at: http://www.sse.com.cn/sseportal/en_us/ps/sczn/sse_indices_cal_and_main_en.pdf, last accessed on 28/01/2024.

[48] Euronext (2015), Euronext 100 Index, Next 150 Index, (2015), available under: https://live.euronext.com/sites/default/files/documentation/index-rules/euronext_100_next_150_index_rules_version_15-01_oct_2015.pdf, last accessed on 28/01/2024.

[49] CME Group (2024), website, at: https://www.cmegroup.com/markets/energy/crude-oil/light-sweet-crude.contractSpecs.html, last accessed on 28/01/2024.

[50] ICE (2020), Benchmark Administration. Statement of Compliance With the Benchmarks Regulation and Independent Assurance, at: https://www.theice.com/publicdocs/Statement_of_Compliance_with_the_EU_Benchamrks_Regulation.pdf, last accessed on 28/01/2024.

[51] V. López, M. Santos, J. Montero, Fuzzy specification in real estate market decision making, Int. J. Comput. Intell. Syst. 3 (1) (2010) 8–20.

[52] Q. Chen, Z. Meng, X. Liu, Q. Jin, R. Su, Decision variants for the automatic determination of optimal feature subset in RF-RFE, Genes. (Basel) 9 (6) (2018) 301, https://doi.org/10.3390/genes9060301.

[53] L. Yang, X.Y. Liu, X. Li, Y. Li, Price prediction of cryptocurrency: an empirical study, in: In Smart Blockchain: Second International Conference, SmartBlock 2019, Birmingham, UK, October 11–13, 2019, Proceedings 2, Springer International Publishing, 2019, pp. 130–139.